Report: Application of Self-Supervised Learning for Wheat Grain Image Classification

---

1. Introduction

This report explores the application of Self-Supervised Learning (SSL) for a wheat grain image classification task. The dataset consists of images belonging to three distinct classes (e.g., different species, disease states, or quality grades). The core challenge in many agricultural domains is the high cost and expertise required to label data. SSL presents a powerful solution by allowing a model to learn robust visual representations directly from unlabeled images, which can then be fine-tuned with a minimal set of labels to achieve high performance on the classification task.

We will analyze three SSL methods—**SimCLR**, **BYOL**, and **Barlow Twins**—justifying their suitability for the wheat dataset and providing a detailed, conceptual explanation of each.

---

2. Selected SSL Methods and Justifications

Our wheat image dataset likely has the following characteristics:

- **High Visual Similarity:** Different classes of wheat grains may look very similar, requiring the model to learn fine-grained, discriminative features.
- **Limited Labeled Data:** A common constraint in agricultural applications.
- **Color and Texture Variance:** Images may have variations in lighting, orientation, and background.

Based on these, we select the following SSL methods:

1. **SimCLR (A Simple Framework for Contrastive Learning of Visual Representations):** Ideal for learning to distinguish between subtle visual features by directly contrasting similar and dissimilar images.
2. **BYOL (Bootstrap Your Own Latent):** Excellent for stabilizing the learning process without needing negative pairs, which can be beneficial when dealing with intra-class variance.
3. **Barlow Twins:** Focuses on reducing redundant information in the learned features, forcing the network to capture semantically meaningful and distinct characteristics of the wheat grains.

---

3. Method 1: SimCLR

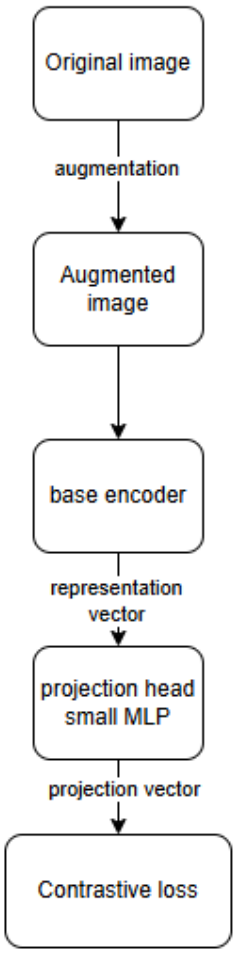
3.1. Intuition

"Teach a model by showing it what is the *same* and what is *different*." SimCLR learns by taking one wheat image, applying two different random transformations (e.g., cropping, color jitter), and then training the model to recognize that these two altered views (called "positive pairs") are similar to each other and different from the altered views of *any other* wheat image in the batch (which are "negative pairs").

3.2. Objective & Diagram

The objective is to maximize the agreement (similarity) between the representations of the two augmented views of the same image while minimizing the agreement with the representations of all other images in the same batch.

The following diagram illustrates the SimCLR framework:



1. An original wheat image is transformed into two correlated views.
2. A base encoder (e.g., a CNN like ResNet) extracts representation vectors.
3. A small neural network (the projection head) maps these representations to a latent space where the contrastive loss is applied.
4. The Normalized Temperature-scaled Cross Entropy Loss (NT-Xent) is calculated

3.3. Positives and Negatives

- **Positives:** The two augmented views (derived from the *same* original wheat image form the positive pair.
- **Negatives:** All other images in the current mini-batch serve as negative examples when calculating the loss for the positive pair

3.4. Why SimCLR fits the Wheat Dataset

SimCLR is perfect for learning to discriminate between visually similar classes. By forcing the model to pull different augmentations of the *same* wheat grain close together in the representation space while pushing it away from images of *different* wheat grains (even if they look similar), it learns to focus on the most salient and invariant features (e.g., grain shape, texture, crease details) that truly define a specific class, ignoring nuisance variations like lighting or orientation.

---

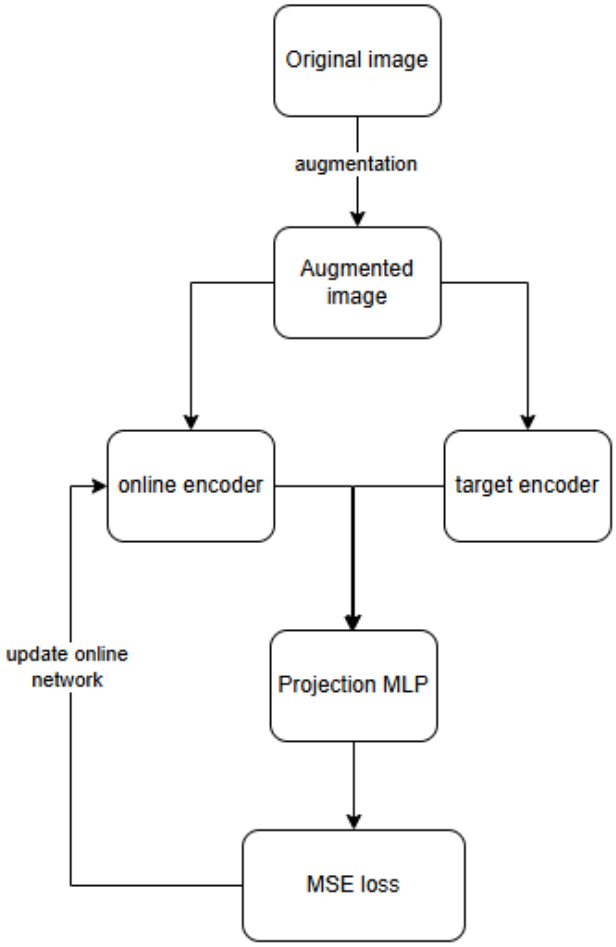
4. Method 2: BYOL (Bootstrap Your Own Latent)

4.1. Intuition

"Learn by predicting yourself, but with a moving target." BYOL uses two neural networks: an **online network** and a **target network**. The online network's goal is to predict the representation of one augmented view of an image produced by the target network from another augmented view of the *same* image. The key is that the target network is not trained by gradient descent; instead, it is a slow-moving, stable average of the online network. This prevents a degenerative solution where both networks output the same constant representation regardless of the input.

4.2. Objective & Diagram

The objective is to minimize the mean squared error between the prediction of the online network and the projection of the target network.



1. Augment views are created from one image.
2. The **online network** (parameters θ) processes one of the augmented image to produce a representation, a projection, and finally a prediction
3. The **target network** (parameters ξ) processes the other augmented image to produce a target projection target(z_y).
4. The MSE loss is computed between the prediction q(z_x) and the target target(z_y).

5. Only the online network is updated via gradient descent.
6. The target network is updated as an Exponential Moving Average (EMA) of the online network

4.3. How it avoids Collapse without Negatives

BYOL cleverly avoids representation collapse (where all inputs map to the same point) without using negative pairs. The "moving target" provided by the target network is crucial. Because the target representation is evolving slowly, the online network cannot simply output a constant value to minimize the loss; it must learn to predict a meaningful, stable representation that works for all augmentations of an image.

4.4. Why BYOL fits the Wheat Dataset

BYOL's stability is a major advantage. For a dataset where different wheat classes might share many features, the constant "push" from negative examples in SimCLR could sometimes be too harsh, potentially making the model ignore useful shared characteristics. BYOL's collaborative, "self-predictive" approach allows it to learn a more robust and balanced representation of the wheat grains, which can lead to better generalization when fine-tuned on the small labeled dataset.

---

5. Method 3: Barlow Twins

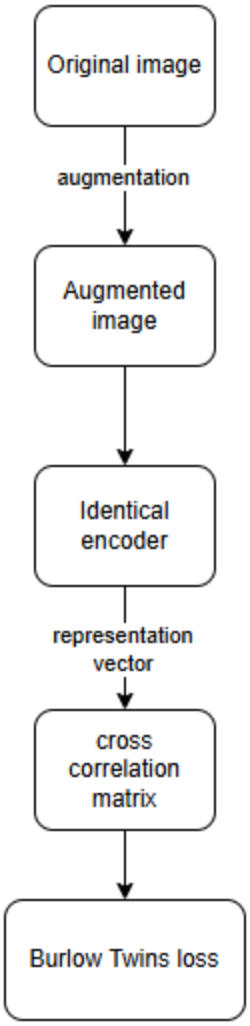
5.1. Intuition

"Learn useful features by making them as statistically independent as possible." Barlow Twins takes two augmented views of an image and passes them through identical networks. The objective is to make the cross-correlation matrix between the output feature vectors as close to the identity matrix as possible. This means:

- Each feature should be similar for the two augmented views (the diagonal should be 1).
- Different features should be uncorrelated with each other (the off-diagonals should be 0).

5.2. Objective & Diagram

The loss function has two parts: an **invariance term** (for the diagonal) and a **redundancy reduction term** (for the off-diagonals).



1. Two augmented views are created.
2. They are fed through identical encoder networks to produce normalized representations
3. The cross-correlation matrix C is computed between two augmented views
4. The Barlow Twins loss is computed. The first term tries to make the diagonal of correlation matrix equal to 1, ensuring each feature is consistent across views. The second term tries to make the off-diagonals 0, decorrelating the features from one another.

5.3. How it works without explicit negatives

Like BYOL, Barlow Twins does not use negative pairs. Instead, it achieves a non-collapsed solution by directly imposing a structure on the representation space itself—specifically, that the features should be statistically independent. This naturally prevents all units from learning the same thing.

5.4. Why Barlow Twins fits the Wheat Dataset

The redundancy reduction objective is powerful for learning highly discriminative features. For wheat grains, many low-level features (e.g., color histograms, edge detectors) might be correlated. Barlow Twins forces the network to find a set of features where each one carries unique, non-overlapping information. This process is excellent for discovering the subtle, high-level features (e.g., specific texture patterns or shape deformations) that are most critical for distinguishing between the three wheat classes.

---

6. Conclusion

We have justified three powerful SSL methods for the wheat grain classification task. **SimCLR** provides a strong, intuitive baseline through contrastive learning. **BYOL** offers a stable, collaborative alternative that avoids potential pitfalls of negative sampling. **Barlow Twins** uses a principled, information-theoretic approach to learn a non-redundant and highly discriminative feature set. An empirical study would involve pre-training a backbone CNN (like ResNet-50) on the unlabeled wheat images using these three methods, followed by fine-tuning a linear classifier on a small labeled subset to compare their final classification accuracy.